

---

**Primary structure of the soybean nodulin-23 gene and potential regulatory elements in the 5'-flanking regions of nodulin and leghemoglobin genes**

---

Vincent P. Mauro, Truyen Nguyen, Panagiotis Katinakis and Desh Pal S. Verma\*

---

Plant Molecular Biology Laboratory, Department of Biology, McGill University, 1205 Docteur Penfield Avenue, Montreal, Quebec, H3A 1B1, Canada

---

Received 27 September 1984; Accepted 3 December 1984

---

**ABSTRACT**

The nodulin-23 gene of soybean is one of the most abundantly transcribed genes induced during symbiosis with *Rhizobium*. Using a plasmid (pNod25) from a nodule cDNA library, we have isolated the nodulin-23 gene from a soybean genomic library. Nucleotide sequence analysis of the cDNA and of the genomic clone indicated that the coding region of this gene is 669 bp long and is interrupted by a single intron of about 530 bp. The deduced protein sequence suggests that nodulin-23 may have a signal sequence. The 5'-flanking sequence of two other nodulin genes, nodulin-24 encoding for a membrane polypeptide and one of the leghemoglobin genes (LbC<sub>3</sub>), were obtained. Comparison of these sequences revealed three conserved regions, one of which, an octanucleotide (GTTTCOCT), has 100% homology. The conserved sequences are arranged in a unique fashion and have a spatial organization with respect to order and position, which may suggest a potential regulatory role in controlling the expression of nodulin and leghemoglobin genes during symbiosis.

**INTRODUCTION**

A number of molecular events occur in the successful infection of leguminous plants by *Rhizobium* species leading to the formation of root nodules. Dinitrogen is reduced inside these specialized structures provided that the essential plant and bacterial genes are expressed in a coordinated manner (see reviews, refs. 1-3). While information on the *Rhizobium* genes is rapidly accumulating, the contribution of the host to this symbiosis is not well understood. At present, only two nodule-specific plant gene products have been characterized. These are leghemoglobin (4,5) and nodulin-35 (6), a subunit of the nodule-specific uricase (7).

A number of other nodule-specific host proteins (nodulins) have been identified (8,9) and the mRNAs coding for these nodulins have been cloned from soybean (10). One of these cDNAs, pNod-25, is a member of the NodA class which represents the second most abundantly transcribed sequences in nodules. In vitro translation of hybrid-selected mRNA using pNod-25 produced two polypeptides of molecular weight 23,500 and 24,500 (10,11). These products are termed nodulin-23.

---

Nodulin-23 is induced early in infection at approximately the same time as leghemoglobin and nodulin-24 (10,11). This induction occurs prior to and is independent of the appearance of nitrogenase activity in nodules (11). These two nodulin genes are not linked to any of the leghemoglobin gene loci (12) but are induced at the same time in response to an apparent common stimulus, the Rhizobium infection. Therefore, it is likely that they share some common regulatory regions as has been observed in other sets of coordinately induced genes (13).

Isolation of the genomic sequence encoding nodulin-23 and the comparison of 5'-flanking sequences of this gene with nodulin-24 and leghemoglobin C3 gene revealed three regions of sequence homology with some unusual, conserved features. A low probability of random occurrence of these sequences may suggest their possible role in regulation of expression of these genes during symbiosis.

### MATERIAL AND METHODS

#### Isolation of the nodulin-23 gene from a soybean genomic library

Approximately 500,000 recombinant bacteriophages were screened using pNod25, a cDNA clone, as probe according to the method of Woo (14) from an AluI/HaeIII partial genomic library of soybean constructed in Charon 4 (15). Two genomic clones were isolated and mapped with the restriction enzymes EcoRI and HindIII. The map of these two phages were found to be identical except for a difference in one fragment. One of the genomic clones GmN7 was chosen for further study. HindIII/EcoRI and EcoRI fragments hybridizing to pNod25 were subcloned into pBR322.

#### DNA sequencing

Nodulin-23 gene with its 5'-end and the 5'-ends of nodulin-24 (16) and leghemoglobin C3 (5) genes were sequenced. Appropriate DNA fragments were digested with various restriction enzymes and cloned in M13 vectors, mp8 or mp9. The clones were propagated in the host JM101 and single-stranded DNA was purified from phage according to the protocol by Amersham Inc. DNA was sequenced primarily using the dideoxy chain termination method (17). Some fragments were sequenced by the method of Maxam and Gilbert (18).

#### Computer analysis

Sequences were assembled using a computer program by Intelligenetics Inc. (Palo Alto, CA).

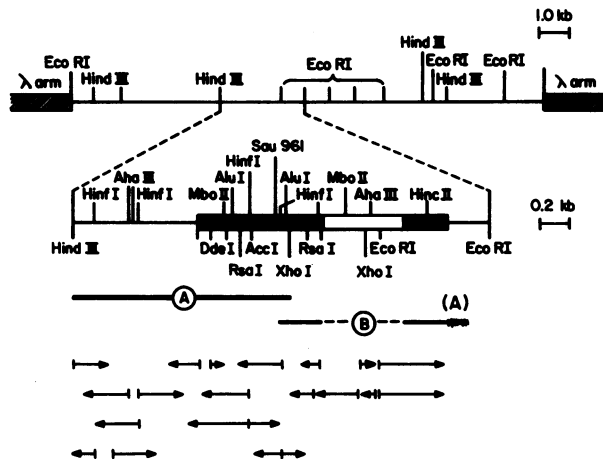


Figure 1. Restriction map and sequencing strategy of a genomic clone (GmN7) containing nodulin-23 gene. The filled regions indicate exons as determined by hybridization studies, *ExoVII* experiments (19) and comparison to cDNA clone pNod25. The open box is an intron. Line A indicates the region used in *ExoVII* experiments to determine the position of any introns in this region (19). Line B indicates the cDNA sequence corresponding to the genomic clone. (A) represents the poly(A) tail on cDNA.

## RESULTS

### Primary structure of nodulin-23 gene

Hybridization of the cDNA clone, pNod25, to *Glycine max* cv. Prize genomic DNA revealed a 7.0 kb *EcoRI* fragment containing the nodulin-23 sequence (10). Two fragments corresponding to this sequence were isolated from a genomic library of soybean. The general organization of one of the recombinant phage (GmN7) is shown in Figure 1. This clone contains single *HindIII* and *EcoRI* hybridizing fragments, the sizes of which are consistent with the results obtained with genomic blots (ref 10 and data not shown). The complete nucleotide sequence of the nodulin-23 gene was determined, using the M13 sequencing strategy shown in Figure 1. The position of the exon was determined by partial sequencing of pNod-25 and from the results of *ExoVII* and S1 mapping experiments (19). The nucleotide sequence is presented in Figure 2.

Sequence analysis of the nodulin-23 gene revealed a coding region of 669 base pairs containing one intron. The first exon is large, coding for 216 amino acids while the second exon codes for only seven amino acids. This gene has several features common to other functional eucaryotic genes: the intron/exon junctions obey the GT/AG boundary rule of Breathnach et al. (20); the 5'

```

-170      -160      -150      -140      -130      -120      -110      -100      -90      -80
AGTAATTAAGTTTAATGATAAAATATATTCTACAGATATATTTCTGTCTCTGGCAACTCGTGAGAATTGAATATTATTATAAGATGAAGGTCGTTACAA
-70      -60      -50      -40      -30      -20      -10      1      10      20
TTTTTTTGAATAAATATTTTATACAAATCTAGATTGTTATATAAATTCACATATTGTATGAGTATAATACATGACACACACCAAACTAGTCTCAAT
      40      50      60      70      79      85      100      110
ATTAAGTAAGGGCTAATTATTAGCGTAGCTAAGTAACCAAGTAATTA ATG GAG AAA ATG AGG GTG ATA GTA ATT ACT GTA TTC
Met Glu Lys Met Arg Val Ile Val Ile Thr Val Phe

      130      145      160      175
CTA TTT ATA GGT GCA GCA ATT GCA GAA GAT GTT GGT ATT GGT CTC CTT AGC GAA GCT GAG GCG TAT GTG TCT CCT
Leu Phe Ile Gly Ala Ala Ile Ala Glu Asp Val Gly Ile Gly Leu Leu Ser Glu Ala Glu Ala Tyr Val Ser Pro

      205      220      235      250
AAG TTA AAA AAG TTC ATC ACA CCT TGC ACT TCG CAT GTT GGT TGC ACA TGC AGT ACT AGT AGT GGA AGT
Lys Leu Lys Lys Phe Ile Thr Pro Cys Thr Ser His Val Gly Glu Thr Cys Ser Thr Thr Ser Ser Gly Ser

      280      295      310      325
GAA GCA TTA ATG CAG AAC CAG GGT GGG TTG CTC TTT GCC TTT CGA TTC TAT GGA GAG ATG CTT GGT AGA CCA TGT
Glu Ala Leu Met Gln Asn Gln Gly Gly Leu Leu Phe Ala Phe Arg Phe Tyr Gly Glu Met Leu Gly Arg Pro Cys

      355      370      385      400
GCC CAA CTT TAT CAA ACA AGT GTT ACT AAC CTT CAA GTT GAA CCT TCT GAA GTA TTT CCA AGC AAG AAT AAT CCA
Ala Gln Leu Tyr Gln Thr Ser Val Thr Asn Leu Gln Val Glu Pro Ser Glu Val Phe Pro Arg Lys Asn Asn Pro

      430      445      460      475
CAA GGT GGA CGT AAG TCC AAA TTA GAT GAC CAT CAA GTT CAA CCC TTA TCA TTT CGA TTA CCA CCA TTT CGA TTA
Gln Gly Gly Arg Lys Ser Lys Leu Asp Asp His Gln Val Gln Pro Leu Ser Phe Arg Leu Pro Pro Phe Arg Leu

      505      520      535      550
CCA CCA ATG CCA AAA CTA GGA CCA ACA AGT CCG ATT ATA AGA ACG ATT CCA TCA CCA CCC ATA GCT CCT CGA GAT
Pro Pro Met Pro Lys Leu Gly Pro Thr Ser Pro Ile Ile Arg Thr Ile Pro Ser Pro Pro Ile Ala Pro Arg Asp

      580      595      610      625
TTG TCA CTC ATT GAG ACT ATA CAA TTA CGA ACT GCC TTG AGA ACC TGT ACT CAT GTC ACT CCA CGA ACT TGT CTC
Leu Ser Leu Ile Glu Thr Ile Gln Leu Arg Thr Ala Leu Arg Thr Cys Thr His Val Thr Ala Arg Thr Cys Leu

      655      670      685      700
ACT GCT CCA AAT GTT GCC ACA TCT GAT TTA GAG GCT TGT CTC ACT CCA TCC ATG AAT CAA TGC ATC TAT CCT GGT
Thr Ala Pro Asn Val Ala Thr Ser Asp Leu Glu Ala Cys Leu Thr Pro Ser Met Asn Gln Cys Ile Tyr Pro Arg

      740      750      760      770      780      790      800
GCA GCT GAA TAT G/GTACGTGTGTTCTTCCTTAATCTCCGTTTTTTTCATAAAAGAAACCATTTTATAAAATATGAAATACGATTAATACATTTTAT
Gly Ala Glu Tyr

      820      830      840      850      860      870      880      890      900
TTTTGCATTTTACTTACAACTTACTTTCATGTATAAACCTGTGAAAAAATAAATTCATCTGACTCTGCATCTATTATTTATACCTTCTAAATGTTTTT

      920      930      940      950      960      970      980      990      1000
AGTTTATACAAATGTGCTATTCATCAATCAATTTTATTCATACAAATACAGGAATAATGCATTCCTTCATGTAAATTTTCGAGTTAGATTTTACATAA

      1020      1030      1040      1050      1060      1070      1080      1090      1100
ATTTTCCATTTTAAAAAATGTTATTTCTCTCTTTTAAATATATGGACAATATAAAATTTGATCGANATTTTACACAAATACATACAGCTAAGTCATTAA

      1120      1130      1140      1150      1160      1170      1180      1190      1200
ATTTATGTGTAATCCCGTGTAATACTACTATGTTTTAGTAGTATAATAATGTCGCTTTTAAATGATTTTCTTTCTTAAATTAATATATTTATCAGTTGTTT

      1220      1230      1240      1250      1270      1280      1290
AGATTTTAAITTAAGTTTTTAAAAATTTATTTTGTGATGCAG/GT AGC CCG CCT ATT AGG GCG TAAATTTATTCAGGCATAAAGGAAATATGTT
Gly Ser Pro Pro Ile Arg Ala ***

      1310      1320      1330      1340      1350      1360      1370      1380      1390
GGAAGTATTGCTAGTAGAAGAATAGCCAGACATGGCTCAACTAGATTTAAGCTAGCTAGCTGTTTTATGATGAAGAGATTGTTATACCTTCAAATTTCC

      1410      1420      1430      1440      1450      1460      1470      1480      1490
AAGTAGCTACTACATGTCATATAAATTAACATTTGAATGTGCAGAACGATATGGTCTCTACTAATATATAAAGCATGCTTATTTTGGTTGATCTAG
      .

```

**Figure 2.** Nucleotide sequence of nodulin-23 and predicted amino acid sequence. Transcription start site (numbered as 1) was determined by S1 mapping (19). The 'TATA' and 'CAAT' boxes corresponding to the major transcription start site are underlined. A minor transcription start site is indicated by a closed triangle. The 'anti-nodulin' transcription start site occurs on the other strand and is indicated by the open triangle. The proposed 'TATA' box for this promoter is overlined. There are three potential polyadenylation signals (underlined) located near the 3' end of the gene. The site of transcription termination is marked with a dot. This site corresponds to the position of polyadenylation in pNod25. Note: there is a 'N' at position 1076 which represents a (less than 10 bp) gap in the intron sequence.

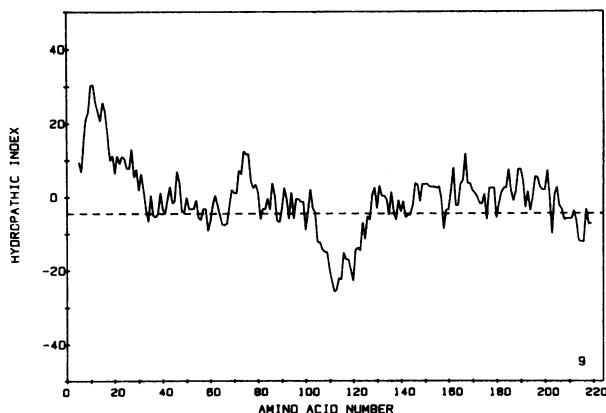


Figure 3. Hydropathic character of proposed nodulin-23 peptide. The hydropathic index is a relative scale in which positive values reflect hydrophobicity and negative values reflect hydrophilicity (40). The window size used in the calculation of the hydropathic profile was 9 amino acids.

region contains 'TATA' and 'CAAT' boxes and there are three putative poly(A) addition signals located 16, 57 and 61 bp from the 3' end of the message (illustrated in Figure 2).

The coding region of this gene starting 80 bp from the major transcription initiation site (19) was chosen as being the only long open reading frame which corresponds to the cDNA sequence (data not shown). In order to assess whether the reading frame resembles that of a real gene or whether it is simply a random open reading frame, we determined the codon usage of nodulin-23. The C-statistic was calculated as being the observed codon usage divided by the expected codon usage of a random sequence. We obtained a value of  $C = 2.095$  which indicates a significant codon bias. Therefore, the reading frame we propose has features of protein coding genes (reviewed in ref. 21). We also calculated the C-statistic for a number of other soybean genes including nodulin-24, nodulin-35, leghemoglobins a, C1, C3, actin and lectin. These results indicate that the degree of codon bias in nodulin-23 is consistent with that of other soybean genes.

#### Structural analysis of the nodulin-23 protein

A putative protein sequence (figure 2), deduced from the DNA sequence is 223 amino acids long and suggests the molecular weight of nodulin-23 to be 24,275. This size is in agreement with one of the hybrid-released translation products using the cDNA clone, pNod-25 (11). Hydropathic analysis (Figure 3) revealed that the amino terminus of the protein is very hydrophobic and con-

A

```

-720      -710      -700      -690      -680      -670      -660      -650      -640
LBC3      AAAATACACTCATATATATATGCCATAAGAACCAACAAAAGTACTATTTAAGAAAAGAAAAAAACCTGCTACATAATTCGAATCTTGAGAT
N23      GAAAGCTATTAAAGAGAAGTGTTAAGAAAAGAGGTTAGCACACCAATAGAAGTATTGAGTTATATTAACCTTTAGATTCTTTTCAAAATGTTTAC
N24      GCAAGCTTAAGTAATATTCTAACCGCAATGACATTAAGGTGATTGGTGAATTAATGAATAAGA

-620      -610      -600      -590      -580      -570      -560      -550      -540
LBC3      TTATTTCTTTATTTTATAAAGGAGAGTTAAAAAAATACAAAAATAAAATAGTGAACATCGTCTAAGCATTTTATATAAGTAGAATTTAAAA
N23      ATTGCATATAGAATTTTATTGACAATCCTTATAACAGTTGCTACTGTTGAAAGACGTTCTTCAAAATTAATACTTAATCATATCTAAAAAT
N24      TTATTTTGATGATAAATATGAATATCATAGATTAATTAATTAACAATTTTACGGTAAATAGTTATTTTCATTGTATAAGGTAGTGACAAATTCATC

-530      -520      -510      -500      -490      -480      -470      -460      -450
LBC3      TATACCATTAATCTGACGGGAACGGATGGATCTTGGCTCAGACCGGAGGAGGAAGACATTTTAAAAATTGGATAAGATCACACTATAAA
N23      CAACAATGTTACAGATAGATTGAATGAGTTAGTTATTTATCTATTGAAAGTAAAGTGTGTAATGTTGATTATAAAACTCGATAAATGA
N24      ATTCAAAGATGAAAAATCAAATTTTAGTCCATAAACTAAAAAAGTGTGACAAATTCATTTATTCATTAACTTCATATATTACGGTTAATGAA

-430      -420      -410      -400      -390      -380      -370      -360      -350
LBC3      GTTCTTCTCCGAGTTTGATATAAAAAAATTTGTTCCCTTTTGGATTATGGATAAAATCTCGTAGTGACATTATTTAAAAAAATAGGGCTC
N23      TTTTGCAGTTAAAAAACTAGAAGATTAAATATAAAATTTGATATTTATATATATTAAGTCTCTTTAAAAATCTTGTAAAAAAGACATTT
N24      AGAGTCTATGTAATATAGAGACAAAAATGATATTTATCCAAAAATATAATTTAATTTTATCTTTCTTTTTTTTACTCGTCATTTTCATAATAT

-340      -330      -320      -310      -300      -290      -280      -270      -260
LBC3      AATTTTTATTAGTATAGTTTGCATAAATTTTAACTTAAAAATAGAGAAAACTGGAAAAGGGCATGTTAAAAAGTGTGATATTAGAAATTTGTC
N23      TTAATAATAAAAAAAGCAACTCTTAATTTTAAAGAAACATCCCTTTGTTAAACCGGAATCTTCCATAATGTAAAAAATTAATGCTTTGATGGAAG
N24      TATAATAAACACTTAAACATAAAAAATCAATATAATTAGACAAACATAAGTAGAATACATTTTGTGTCATTTTTCACAAAATCGAAATTAATAA

-250      -240      -230      -220      -210      -200      -190      -180      -170
LBC3      GGATATATTAATTTTATTTTATATGGAACTAAAAAAATATATATTAATAATTTTAAATTCAGAAATAATCTTAAATTTATTTTACTGAAA
N23      TTTTAAATTTGTTCTATCCAACTCAAGGGTTGTAATATTTTTTTTATCATTTATATGTTGTAATATGAATGGACATAGTAATAGTTTAA
N24      TTTTAAATTTTATTTTAAAAATGAATTTGTTGACATTTCTACATTTTCAAAGAAAAGATAACTATTAGAGTATTTTAACTAATTTT

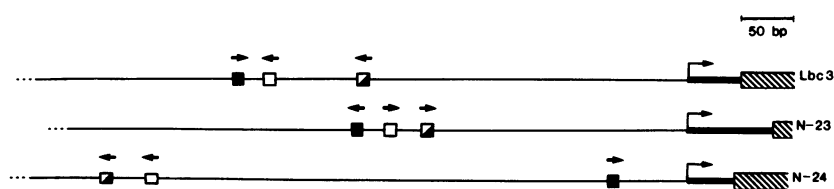
-150      -140      -130      -120      -110      -100      -90      -80      -70
LBC3      TGAGTTGATTTAAGTTTGAAGAGATGATGTCCTTCCACCATACCAATTGATCACCCTCCCTCAACAAGCCAAGAGAGACATAAGTTTATTA
N23      TGATAAAATATATTCTACAGATATATTCTTGCTCTTGGCAACTCGTGAGAATTGAATATATTATAAAGATGAAAGGTCGTTTACAATTTTTTTT
N24      GTGCTCTTACATAATAATAAATCAGTTAGTAACTTTGAATTTCTGAATTAATATGTTACGGGCCAGGTGAAGAAAGAGGAGTTTCCCTATTT

-60      -50      -40      -30      -20      -10      1      10      20      30
LBC3      GTTATCTGATCACTCTTCAAGCCTTCTATATAAAATAGTATTGGATGTGAAGTTGTTGCATAAATCTGCATTGAACAATTAATAGAAATAACAG
N23      AGAATAAATATTTATACAAATCTAGATTTGTTATAAAATTCACATATTGTATGAGTATAATACATGAGCACACCAACAACTAGTCTCAAA
N24      CTCCTACTCCAACCTCTTATATAGAGTATATATCCCAAAATTTTCTCATCTTTTGTACTAAACAACTCGATCTGTTGTAATTTTATTAGT

40      50      60      70      80
LBC3      AAAAGTAGAAAAAATATG
N23      TAAGTAAGGGGCTAATTATTAGCGCTAGCTAAGTAACCAAGTAATTAATG
N24      ACGTATTGAAAAATG

```

B



**Figure 4.** (a) Sequence of the 5'-flanking sequence of Lbc<sub>3</sub>, nodulin-23 and nodulin-24. The 5'-sequence of Lbc<sub>3</sub> was obtained using the genomic clone pLb11-3.7 (5). The 5'-sequence of nodulin-24 (N24) was obtained from the genomic clone GmN-24 (16). The 5'-flanking regions of nodulin-23 (N23) was sequenced as shown in Figure 1. Consensus sequence 'A' (see Table 1) is underlined by a thick line. Consensus sequence 'B' is underlined by a thin line. Consensus sequence 'C' is underlined by a dotted line. Numbering starts at the major transcription start site for each gene. (b) Illustrations of the consensus sequences in the 5'-flanking regions of nodulin-23, nodulin-24 and Lbc<sub>3</sub> genes. The solid box represents consensus sequence 'A'. The blank box represents consensus sequence 'B'. The diagonally shaded box represents consensus sequence 'C'. The relative orientations of the consensus sequences are marked by arrows above the boxes. The curved arrows indicate the transcription start sites. The thick line is untranslated 5' region. The hatched area is the coding region.

TABLE 1. 5' CONSENSUS SEQUENCES IN NODULIN AND LEGHEMOGLOBIN GENES

Consensus 'A'	TAAAAAAATTGTTTCCCTATT	Homology to Consensus		Probability of Random Occurrence
		8 bp	22 bp	
	8 bp consensus			less than $2.5 \times 10^{-5}$
	22 bp consensus			less than $9.5 \times 10^{-7}$
Lbc <sub>3</sub>	..... TAAAAAAATTGTTTCCCTTTT	100%	95%	
inv N-23	... .. TAAGCCAAATTGTTTCCCTACA	100%	77%	
N-24	. . . . . AAGAAAGAGGAGTTTCCCTATT	100%	73%	
Consensus 'B'	TATAAGGTAGTGACAAATTAATA	23 bp		less than $8.9 \times 10^{-5}$
Lbc <sub>3</sub>	.. .. AATCTCGTAGTGACATTATATTA	65%		
inv N-23	. . . . . TTGAAGGTAGTTGTAATTAAAA	70%		
N-24	..... TATAAGGTAGTGACAAATTCATC	91%		
Consensus 'C'	TCTGGGAAA	9 bp		less than $4 \times 10^{-2}$ for Lbc <sub>3</sub> and inv N-23
	.....			
Lbc <sub>2</sub>	TCTGGGAAA	100%		
Lbc <sub>3</sub>	..... TCTGGGAAA	88%		
inv N-23	..... GTTGGGAAA	78%		
N-24	. . . . . TTAGGGTAA	67%		

tains one strongly hydrophilic region near the center of the molecule. The amino terminus of the peptide has features of a signal peptide (22). It contains three charged residues within the first five followed by 15 consecutive hydrophobic residues and is terminated with a polar glutamine residue. A potential cleavage site could occur at the glycine or serine residues, located five and eight amino acids downstream of the glutamine residue (22). It is not known where nodulin-23 is transported in the nodule cell, but localization studies are in progress to determine its intracellular site.

#### 5'-flanking regions of nodulin and leghemoglobin genes

The 5'-flanking sequences for nodulin-23, nodulin-24 and leghemoglobin C3

genes were obtained and are shown in Figure 4a. Homology searches of these sequences revealed three major conserved regions. The calculated probabilities of the occurrence of such sequences are shown in Table 1. The probability calculation takes into account the length of the consensus sequence, the number of matches each sequence has to this consensus and the location relative to the start of the gene where these sequences are found. The most conserved sequence found (consensus 'A') is an octanucleotide 'GTTTCCT'. This sequence is inverted in the nodulin-23 gene. Occurrence of this sequence by chance is unlikely since even a 22 nucleotide consensus sequence derived from this region shows between 73% to 95% homology for the three genes. The chance occurrence of a sequence 22 bases long with this homology in the location found was calculated as being less than  $9.5 \times 10^{-7}$  (based on the calculations in ref. 23).

A second area of homology, containing the core 'GGTAGTG' (consensus 'B'), was found to have a probability of occurrence at less than  $9 \times 10^{-5}$ . This sequence was also inverted in the nodulin-23 gene. A third area of homology, with the core 'TCTGGGAAA' (consensus 'C'), was found but its occurrence by chance could not be ruled out when the three sequences were compared. However, its occurrence between the Lbc3 and nodulin-23 does not appear to be by chance. This sequence is also present in another leghemoglobin, Lbc<sub>2</sub> gene (data not shown). Like the other two consensus sequences, this one is also inverted in nodulin-23. The location and orientation of these three consensus sequences is shown in Figure 4b. Consensus 'C' shows a diad symmetry to the consensus 'A'. The probability that these three elements together would be located by chance in that position in the Lbc3 and nodulin-23 genes is unlikely (P is less than  $4 \times 10^{-11}$ ).

### DISCUSSION

Early studies in our laboratory have revealed several interesting features of the nodulin-23 gene: it is transcribed at a high level (10), and hence, may have an important role in symbiotic nitrogen fixation; its transcription occurs early in infection prior to nitrogen fixation along with leghemoglobin and a few other nodulins which may indicate a coordinate regulation of expression and this gene has homology with nodulin-44, another abundantly transcribed sequence in soybean nodules (11). We determined the primary structure of the soybean nodulin-23 gene. The hybrid-released translation of pNod25 (10,11) showed two products (23,500 and 24,500 MW), one of which is represented by this gene. The other product could be either due to processing or it is coded by another gene.

We have isolated a recombinant phage carrying a nodulin-23 sequence with an almost identical organization to that of Gmn7 (unpublished data).

Promoter analysis (*in vitro*) of nodulin-23 gene revealed transcripts starting from two sites 23 bp apart (19). One of these is an animal-like promoter sequence whereas the other is plant-like (24). *In vitro* transcription studies also revealed an 'anti-nodulin' promoter in this gene. This initiates transcription in the direction opposite to that of the sense-strand. The 5'-ends of the nodulin and anti-nodulin messages are complementary to each other for about 90 nucleotides, suggesting a possible regulatory role for the anti-nodulin promoter (see ref.26). In addition, this gene contains three randomly arranged, procaryotic type promoters in the eucaryotic promoter regions. These promoters function when placed in *E. coli* (see ref. 19). The presence of procaryotic features on this gene suggest that a *Rhizobium* interaction may be involved in the regulation of this nodulin gene.

Comparison of the 5'-flanking sequence of nodulin-23 with that of two other physically unlinked genes induced at the same time revealed features which may be involved in the independent regulation of these genes. The consensus regions 'A' and 'C' (as shown in Table 1 and illustrated in Figure 4b) are always inverted and complementary to each other and arranged so as to flank region 'B'. The orientation of the three consensus sequences in LbC3 is the same as in nodulin-24. However, the orientation of all three sequences are found to be inverted in nodulin-23 with respect to those from the other two genes. Notwithstanding orientation, nodulin-23 and LbC3 have a similar spatial organization of consensus sequences, both with respect to order and position (see Figure 4b). The probability of this occurring by chance is indeed very low ( $4 \times 10^{-11}$ ). Computer search of four soybean genes including a non-induced gene with its 5' region failed to reveal the presence of these sequences in any location.

Transcriptional regulation of coordinately induced genes via *cis*-regulatory control was first elucidated by Britten and Davidson (27). Since then a number of coordinately-regulated genes have been shown to share short repeated elements in the 5'-flanking regions. Functional assays of four genes involved in amino acid synthesis in yeast (28) and one of the *Drosophila* heat shock genes, hsp70 (29), showed that the conserved 5' element is sufficient to endow the adjacent gene with the ability to respond to inductive control. In the case of heat shock protein, hsp82, a region of diad symmetry has been demonstrated to bind specifically to an activating protein factor (30). A number of other examples support the hypothesis that short, 5'-conserved sequences are involved

in coordinate regulation. These include the silkworm middle and late chorion genes (31), the human glucocorticoid-repressed proopiomelanocortin gene, the glucocorticoid-induced mouse mammary tumor virus and the rat growth hormone (32-34). Three steroid-induced egg white proteins in chicken (35), the three chains of fibrinogen (36), the immunoglobulin genes (37) and the B-globin genes (38), all contain conserved, putative regulatory sequences located 5' to the coding region. Furthermore, the microsymbiont (*Rhizobium*) of root-nodule has been shown to share common promoter sequences in some of the symbiotic genes (39). This may also reflect in some genes of the macrosymbiont (host plant). The functional analysis of these consensus sequences in nodulin genes must await successful transfer, regeneration, segregation and induction of these genes in the second generation of plant following the association with a suitable microsymbiont.

### ACKNOWLEDGEMENTS

This work was supported by research grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Ministry of Education, Quebec. We wish to thank Sui-Lam Wong for his information concerning the *ExoVII* and *S1* nuclease mapping experiments, Eric Olson for his critical comments on the manuscript, Victoria Foster for her excellent technical assistance, Joel Delisle for his computer analysis and Yvette Mark for her assistance in preparing this manuscript. VM was supported by an NSERC post-graduate fellowship.

\*To whom correspondence should be addressed

### REFERENCES

1. Beringer, J.E., Brewin, N.J., Johnson, A.W.B., Schulman, H.M. and Hopwood, D.A. (1979) *Proc. Roy. Soc. Lond. B* 204, 219-233.
2. Verma, D.P.S. and Long, S. (1983) *Int. Rev. Cytol.* 14, 211-245.
3. Verma, D.P.S. and Nadler, S. (1984) In *Genes Involved in Microbe-Plant Interactions*, D.P.S. Verma and T. Hohn, eds. (Wein New York: Springer-Verlag), pp. 58-84.
4. Baulcombe, D. and Verma, D.P.S. (1978) *Nuc. Acids Res.* 5, 4141-4153.
5. Brisson, N. and Verma, D.P.S. (1982) *Proc. Natl. Acad. Sci. USA* 79, 4055-4059.
6. Legocki, R.P. and Verma, D.P.S. (1979) *Science* 205, 190-193.
7. Bergmann, H., Preddie, E. and Verma, D.P.S. (1983) *The EMBO Journal* 2, 2333-2339.
8. Legocki, R.P. and Verma, D.P.S. (1980) *Cell* 20, 153-163.
9. Bisseling, T., Been, C., Klugkist, J. and van Kammen, A. and Nadler, K. (1983) *The EMBO Journal* 2, 961-966.
10. Fuller, F., Kunstner, P.W., Nguyen, T. and Verma, D.P.S. (1983) *Proc. Natl. Acad. Sci. USA* 80, 2594-2598.
11. Fuller, F. and Verma, D.P.S. (1984) *Plant Mol. Biol.* 3, 21-28.

12. Lee, J., Brown, G.G., Verma, D.P.S. (1983) Nucl. Acid Res. 11, 5541-5553.
13. Davidson, E.H., Jacobs, H.T. and Britten, R.J. (1983) Nature 301, 468-470.
14. Woo, S.L.C. (1979) Methods Enzymol. 68, 389-395.
15. Fisher, R.L. and Goldberg, R.B. (1982) Cell 29, 651-660.
16. Katinakis, P. and Verma, D.P.S. (Proc. Natl. Acad. Sci. submitted).
17. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463-5467.
18. Maxam, A.M. and Gilbert, W. (1980) Methods Enzymol. 65, 499-560.
19. Wong, S.L. and Verma, D.P.S. (1984) J. Biol. Chem. (submitted).
20. Breathnach, R., Benoist, C., O'Hare, K., Gannon, F. and Chambon, P. (1978) Proc. Natl. Acad. Sci. USA 75, 4853-4857.
21. Shepherd, J.C.W. (1982) Cold Spring Harbor Symposia on Quantitative Biology. Vol XLVII, Structures in DNA: 1099-1108.
22. Watson, M.E.E. (1984) Nuc. Acids Res. 12, 5145-5164.
23. Dykes, G., Bambara, R., Mariani, K. and Wu, R. (1975) Nucl. Acid Res. 2, 327-345.
24. Messing, J., Geraghty, D., Heidecker, G., Hu, N., Kridl, J. and Rubenstein, I. (1983) In Genetic Engineering of Plants, T. Kasuge, C.P. Meredith and A. Hollaender, eds. (New York: Plenum Press), pp. 211-227.
25. Izant, J.G. and Weintraub, H. (1984) Cell 36, 1007-1015.
26. Marx, J.L. (1984) Science 225: 819.
27. Britten, R.J. and Davidson, E.H. (1969) Science 165, 349-357.
28. Donahue, T.F., Davis, R.S., Lucchini, G. and Fink, G.R. (1983) Cell 32, 89-98.
29. Pelham, H.R.B. (1982) Cell 30, 517-528.
30. Wu, C. (1984) Nature 311, 81-84.
31. Jones, C.W. and Kafatos, F.C. (1980) Cell 22, 855-867.
32. Cochet, M., Chang, A.C.Y. and Cohen, S.N. (1982) Nature 297, 335-339.
33. Barta, A., Richards, R.I., Baxter, J.D. and Shine, J. (1981) Proc. Natl. Acad. Sci. USA 78, 4867-4871.
34. Donehower, L.A., Huang, A.L. and Hager, G.L. (1981) J. Virol. 37, 226.
35. Grez, M., Land, K.G. and Sahutz, G. (1981) Cell 25, 743-752.
36. Fowlkes, D.M., Mullis, N.T., Comeau, C.M. and Crabtree, G.R. (1984) Proc. Natl. Acad. Sci. USA 81, 2313-2316.
37. Parslow, T.G., Blair, D.L., Murphy, W.J. and Granner, D.K. (1984) Proc. Natl. Acad. Sci. USA 81, 2650-2654.
38. Moschonas, N., de Boer, E. and Flavell, R.A. (1982) Nucl. Acid Res. 10, 2109-2120.
39. Better, M., Lewis, B., Corbin, D., Ditta, G. and Helinski, D.R. (1983) Cell 35, 479-485.
40. Kyte, K. and Doolittle, R.F. (1982) J. Mol. Biol. 157, 105-132.